LTJ_31_1_Language_Samples

[Start of recorded material]

Glenn Fulcher: From the University of Leicester in the United Kingdom, this is Glenn Fulcher with another issue of Language Testing Bites. How large a language sample do we need in order to draw reliable conclusions about what we wish to assess. This question has been asked by every writing and speaking test designer in our field and yet there are very few empirical studies that look systematically at sample domain, congrual, that's in the literature.

In issue 31.1 of Language Testing we are delighted to publish a paper by Jodi Tommerdahl and Cynthia Kilpatrick from the University of Texas at Arlington on the reliability of morphological analyses in language samples. They pose the question, how large a language sample is required for it to be representative of a child's ability in morphology and syntax and their research has clear implications for the assessment of general language knowledge of young learners.

Welcome to Language Testing Bites and thank you for agreeing to talk to our readers and listeners about your research.

Jodi Tommerdahl: Thank you for having us and Dr Kilpatrick and I are both happy to be here.

Glenn Fulcher: Perhaps we can just set the scene for listeners to the podcast who might not be familiar with the field in which you've been working. Can you explain for us why researchers use spontaneous language samples to assess the language of children and how these language samples are collected? I am particularly interested in this because if you say it is spontaneous, presumably you don't use carefully crafted tasks.

Jodi Tommerdahl: Well, first, let me set up what we mean by spontaneous samples. My training in the field of clinical linguistics means that we're looking at extensively child language and very often we're looking at children who have difficulties learning language. Their acquisition patterns aren't normal so we need a very clear picture of the language of each child that we see. So when we're looking at children's language there is really two main categories of language that we're looking at, one is elicited language and the other is spontaneous. So, elicited language means that there is a certain task that we ask the child to do. It could be that they are retelling a story that we read to them but they respond to specific stimuli, something like a test. It could be the description of a single picture or describing a series of pictures, anything of that nature, but its utility is that it is really getting at a certain element of language that we want to look at. So most tests that identify language impairment are set up in some form of test, they are eliciting something specific in language. For example, one form of elicitation might be to look at the past tense that a child can produce. Now on the other hand there are spontaneous language samples and that is really just what it sounds like more or less. It is natural everyday language in a normal context and there are some advantages of spontaneous samples over elicited samples. When you use spontaneous samples the children aren't in any kind of test taking situation so they are not nervous, and there is no problem with anything like attention or motivation but especially the main advantage is that it shows natural, real life language that the children use on a day to day basis. Then we get to the question of know do we actually get these samples and different researchers go about this in different ways. Some researchers send recording equipment home with parents. They might ask the children to have microphones around them at meal time or during a certain half hour every day and sometimes it is not the best quality. You never know exactly how it is going to be placed in the home, you don't know how much freedom the children are going to have to run around and if the microphones will pick up what they are saying, also you can't even be sure who is going to be in the home, there might be siblings or there might be visitors or there might be grandparents, so you don't always get the best quality recordings.

Other researchers record in the clinic with the professional interacting with the child and that can be very good quality-wise but it is not always the most natural, as the child might not be using language with the clinician in the way that they normally would with their family. In our study we were very fortunate to have the best of both worlds. This study was carried out initially with funding from Birmingham University where they had a flexible learning room and that means that we had a large space that we could set up, toys, we had hidden microphones, hidden cameras and what we did was each parent and child together came in and just had a normal play situation for about 40 minutes while they were recorded and because of that we had great quality for both the video and the audio all in a very naturalistic play setting.

So most of what we know about child language actually comes from spontaneous samples, meaning it comes from observing children communicate, and ideally when you are diagnosing language impairment, a combination of elicited and spontaneous language is the best.

Glenn Fulcher: So once you have collected these samples, what kinds of features do researchers look for? I mean what kind of analyses do they conduct and why?

Jodi Tommerdahl: That question has a very broad range of possible answers and I will try to divide it into two main parts with the first part of it being about using spontaneous samples in a clinical setting. In clinic we might look at a child's language, for example, to examine their morphology and syntax if we think the child may have specific language impairment or SLI. For other problems such as looking for semantic pragmatic disorder a spontaneous sample could be analysed in terms of vocabulary size or vocabulary use. It might even be used to look at the sound system of language, so looking at the phrenology of an individual child. It can also be used for very general measures such as MLU or mean length of utterance or things like the type/token ratio.

My own personal entry into using spontaneous samples really began with my using the LARSP, which stands for the Languages Assessment Remediation and Screening Procedure. It is a linguistic profile developed by Crystal, Fletcher and Garman and it is often used in clinic and it is based on spontaneous samples and the way that it works is we analyse sentences for their morphology and syntax. We look at each sentence individually and whenever a feature occurs we mark it on the LARSP chart and at the end of that that gives us a language age that we can label the child as being at that age as far as morpho-syntax is which we then compare to their chronological age and that gives us a very detailed profile of a child's language and even more useful as sometimes we can see the holes in language, things that aren't being used that we might expect to be used at a certain age.

Now outside of the clinic researchers also use spontaneous samples for a variety of different things. The first thing that comes to mind would be first language acquisition. Researchers ask questions like, 'at what age do certain linguistic traits first develop'? That could be the first use of certain sounds, of specific words, even at what age do children produce their first word and when do they start making two combinations, what is the composition of early vocabulary – all of these questions can be asked.

Really almost everything we know about the early stages of language comes from looking at spontaneous language. In the same vein you could use spontaneous samples for looking at second language acquisition like determining whether second language learners learn a language in the same order as first language learners. You could use spontaneous samples to look at language proficiency in a whole range of different ways through vocabulary, grammatical errors and what have you.

Now we could even use spontaneous samples outside of this area when we're not looking at children. For example, spontaneous samples have often been used for looking at [avasa s.l. 0:08:28.4] which is language loss due to brain damage. This often occurs after strokes in older adults. It can also be used for looking at changes in language that are associated with dementia and so overall spontaneous language samples can be used as potential predictors for a very wide range of problems and to provide us with general insight about language development for both typical and atypical language learners.

Glenn Fulcher: Thanks for making that clear. I can see why reliability is going to be an important issue in this context. Now in your article you state and I quote 'that the temporal reliability of spontaneous samples is important'. Now what do you mean by temporal reliability and how is this estimated?

Cynthia Kilpatrick: I'll take that one if you would like, Jodi. When we talk about temporal reliability we're really just looking at the question of whether samples taken at different times produce consistent results and we're using the terminology 'temporal reliability' to distinguish somewhat from other types of reliability such as inner writer reliability which has been studied quite a bit in terms of sampling. When we're thinking about temporal reliability our interest in this question is how reliable two samples are when they are taken on different days within one week of each other and we're controlling for all types of other factors. So, for instance previous research by Davidson in 2000 shows that factors such as time of day and type of task affects second language proficiency scores and so we're trying to keep those types of things very consistent with each other and really just see if taking the two samples on different days makes a difference in the types of results that we get.

In order to estimate reliability in this way we go back to the idea of classical test theory and the statistical concept of reliability which takes into account the observed score of our participant but also possible measurement or error that may have arisen and the actual test that we run is called the inter-class correlation coefficient or the ICC which measures absolutely agreement between two samples. And when we run an ICC in our design we're looking at a test re-test design and we get a correlation between zero and one to look at agreement between the two samples and when we have a correlation at zero it means there is no agreement. Those two examples are nothing alike.

On the other hand when we have a correlation of one we show complete agreement. Those two samples are virtually identical. Clearly if we're looking at two different samples of language we're never going to get one because otherwise they would just be identical samples. So in our study we’re looking at a correlation of .6 or higher as what we're looking at to see if we can really claim that these samples are consistent with each other. Another aspect of reliability that we're looking at in this study is related to the length of a sample and how long it has to be in order to be reliable. So previous research has looked at a lot of… you might call them more global factors like MLU, the [unintelligible 0:11:41.5]. They've looked at total numbers of words, they've looked at type token ration and different things like that and they've shown that those types of things are quite reliable. But our interest here is in looking at very specific morphosyntactic structures and seeing how consistently they are used between two different samples taken on two different days, but within one week of each other in all instances.

Glenn Fulcher: I guess one of the reasons why this is so important is really a validity issue. If someone is going to give a score for syntax or control of morphology on some kind of scale, it is assumed that there is some direct relationship between the score and the child's language development and this assumes that in this context there is a very real link between the scale, the score and an acquisitional process. This has always been one of the Holy Grails of language testing, so can you tell us a little more about this assumption and how we know it is accurate?

Cynthia Kilpatrick: This is a really great question and this question of validity is actually what led us to this research originally, so many times when we as linguists find a language sample we analyse that language sample to see what's in it and see what the speaker is doing, to see what their morphology, their syntax or phenology looks like and we draw conclusions about the state of their language based on perhaps a single language sample. And so the question that arises with that is how valid is that. Can we really draw a conclusion about the state of a person's linguistic knowledge on the basis of a single sample? And the reason that is a touch question is because we simply don't know if that language sample is consistent if it really truly represents what that person is capable of in terms of their linguistic knowledge, so that is what leads to a study like this where we want to find out how consistent different samples from the same person are so that we can actually draw conclusions about how valid a single sample is as a picture of an individuals linguistic knowledge.

So in many ways we might think about reliability as a necessary precursor to validity. If a test is not going to produce the same results on multiple administrations or in other words if it is not reliable. It is actually very difficult for us to draw any conclusions that link the score on that test to a person's language development. On the other hand if we find that people consistently use language in the same ways in different samples then we can draw the conclusion that the score that we're giving them is related to the process of acquisition in some way but it might not always be in the way that we have expected. So, for instance, in the research that we're presenting here we expected to find higher reliability for targets that we acquired earlier or that are known to be acquired earlier, but in our study that wasn't actually want we found and that is actually leading us to investigate this question further into expand this project to add in more children, add in more samples, to add in a greater number of samples from each child to look at other factors more carefully such as age and gender and things like that and really be able to look at where reliability is emerging so that we know in what ways we can rapidly use our language samples to draw conclusions about linguistic knowledge.

Glenn Fulcher: That's really interesting and very important for our field. Can I ask you to very briefly now tell our listeners about the study that you are reporting in your article in this month's issue of Language Testing, your findings and their significance?

Cynthia Kilpatrick: We are reporting here on a study that examines the reliability of language samples in children using a test, re test methodology where we're looking at nested samples of various lengths, 50, 100, 150 and 200 utterances in samples collected within one week of each other. In these samples we examined five different morphosyntactic categories. Four of these we selected based on Brown's 1973 table of morpheme acquisition. Two of them were considered early acquired, those were plural S and I and G and two of them are late acquired and that is genitive and the copula. Then our fifth morpheme that we looked at is actually a larger category that we termed multi verb utterances and we chose this category because it is clearly morphosyntactic in nature and it is expected to have quite a high frequency and in addition it shows a degree of complexity in the child language that we might not see with a couple of other morphosyntactic categories. In a nutshell our results showed that overall the highest frequency items of multi verb and copula were the most reliable and both of them reached a correlation of .6 or above of by 100 utterances. From these results we draw three major findings. First of all we see that 200 utterances is not significantly better than 100 utterances in terms of reliability of morphosynax within a spontaneous language sample, so 100 utterances might be the optimal sample size since we see no significant increase beyond that. This is actually quite importance because clinicians and linguists and people who work with language samples spend a lot of time recording and transcribing language samples and you want the shortest sample possible to give you the best result you can get or the most reliable result you can get. In our study we're looking at about 100 utterances for that.

A second major finding in this study is related to non-use of particular samples in the study. So some structures were not used at all in one sample by a child but were used in the other sample by the child, and this was actually quite surprising to us, so we saw some children who, for instance, didn't use the genitive at all in their first sample but they used it in their second or conversely they used it in their first sample but they didn't use it in their second one and this was actually quite interesting because it indicates to us that non use of low frequency items doesn't necessarily indicate non mastery, and so this is actually very important possibly in clinical work where we might say you have a spontaneous sample that a clinician has looked at and they might say, oh I don't really see the genitive or I don't see many uses of the past tense and then they can use that to be something that they target specifically within elicited samples with that particular child.

Finally our third major finding is this study is that reliability may not reflect acquisitional stages in the way that we originally thought. So morphemes that are supposed to be early acquired such as plural S and I and G were not actually more reliable than later acquired morphemes like the genitive and the copula and we expected that in some degree forms that were more solidly acquired or that children had acquired over a longer period of time might have had a higher reliability but this isn't actually what we found and this actually is leading us into other work that is looking at how this plays out with different ages and additional samples.

Glenn Fulcher: Before we bring this podcast to an end, I would like to ask you a broader question about the implication of this kind of research to diagnostic assessment for children. What do you think are the implications for the development of diagnostic tests with the specific purpose of discovering language weaknesses or problems and using that information to create individualised learning packages?

Jodi Tommerdahl: Now that one is really the big question and here you are hitting on our main goal. At the moment language impairment is often undiagnosed and when it is diagnosed it is diagnosed later than we want it to be, so precisely what we want to do is to develop a database of linguistic norms of young children so that we and other researchers can use them to compare the language of children at risk with the language of typically developing children. If we can find meaningful differences in typical and atypical children language at age two instead of at age six we're just that much closer to developing more appropriate treatments for those kids and getting them back on track both in the linguistic sense but also educationally. So our plan is to take this research to its next step by attracting funding to either develop or at least contribute to an enormous database of child language that would be freely available to researchers who want to use it recording. Transcribing children's language is a hard job and it takes an enormous amount of time and effort.

However, we have already had four great influences that have inspired us to believe that it is possible to do. The first one is the LARSP assessment that I talked about earlier. It relies on spontaneous samples to assign a language age to an individual based on their language production and then compare that to their chronological age. It was a brilliant idea. Number two was the Wisconsin study by Leadholm and Miller in 1992. They created a database of language from 266 children aged three to 13 and they used frequency counts of different words and grammatical types. They didn't look at reliability but the frequency counts did a great job. They too developed this to better diagnose language impairment and the project shows amazing ambition but we think it should be bigger and even more comprehensive. The third inspiration is McWhinney and Snow's Childes database. This is probably the most influential out of the four I'm going to talk about and this database shows child language from around the world. It shows both clinical cases and normal language and it gives the opportunity for researchers to work on child language without that tedious task of recruiting and recording and transcribing which, honestly, is probably prohibitive for most research labs. So it has really changed the game as far as what child language researchers are able to do. And the last one I will mention is the work on automatic tagging that has been developed by researchers such a Kristoff Paris in Paris and the application of this tagging his collaborators such as Marie-Thérèse Lemonde also in Paris and Cristal Meyer in Belgium with their innovative method in using spontaneous language samples have really inspired us to carry their work further. So following in these footsteps the research possibilities and the potential outcomes are in our eyes limitless.

Glenn Fulcher: Many thanks for telling us about your research and for your insights into these wider issues of assessment. Your article had added greatly to the diversity of assessment interests that are represented in the journal and I am sure that our readers will not only enjoy hearing about what you do but will be inspired in their own research with assessing children. Thank you very much.

Jodi Tommerdahl: Thanks for letting us talk about what we love doing and have us back, there's a lot more to come.

Glenn Fulcher: Thank you for listening to this issue of Language Testing Bites. Language Testing Bites is a production of the journal 'Language Testing' from Sage Publications. You can subscribe to language testing bites through i-Tunes or you can download future issues from ltj.sagepub.com or from languagetesting.info. So until next time we hope you enjoy the current issue of language testing.

[End of recorded material]